1 Краткие теоретические сведения

К дискретным источникам информации относят источники с конечным алфавитом K. Примерами таких источников могут служить тексты на различных языках, телеграммы, e-mail и sms-сообщения, любые файлы данных.

С точки зрения теории информации источник характеризуется степенью неопределённости относительно выдаваемого им сообщения. Количественно эта неопределённость измеряется через **информационные характеристики** данного источника.

1.1 Информационные характеристики дискретного источника

Основной информационной характеристикой дискретного источника является его **энтропия**: среднее количество информации, приходящееся на один символ источника.

$$H(A) = \overline{I(a_i)} = -\sum_{i=0}^{K-1} p(a_i) \log_2 p(a_i).$$

Здесь $I(a_i) = -\log_2 p(a_i)$ — **информация**, содержащаяся в символе a_i, $p(a_i)$ — вероятность его появления, A — множество символов a_i (алфавит источника).

Энтропия измеряется в битах на символ источника: [бит/симв].

Максимально возможное значение энтропии (максимально возможное среднее количество информации на символ) достигается при равновероятном выборе символов источником:

$$H_{\max}(A) = -\sum_{i=0}^{K-1} \frac{1}{K} \log_2 \frac{1}{K} = \log_2 K.$$

Для источника с **памятью** (соседние символы сообщения зависимы) вводится понятие **условной энтропии**:

$$H(A|A') = -\sum_{i=0}^{K-1} \sum_{j=0}^{K-1} p(a_i, a_j) \log_2 p(a_i|a_j).$$

Здесь $p(a_i, a_j)$ — вероятность совместного появления символов a_i и a_j, $p(a_i|a_j)$ — вероятность появления символа a_i при условии, что до него появился символ a_j (условная вероятность символа a_i), A' — множество символов источника на предыдущем шаге (a_j), A — на текущем шаге (a_i).

Условная энтропия не превышает безусловной

$$H(A|A') \leqslant H(A),$$

поскольку всякая дополнительная зависимость не увеличивает (а разве только уменьшает) количество информации.

Избыточность источника

$$\rho_{\text{и}} = 1 - \frac{H(A)}{H_{\max}(A)} = 1 - \frac{n_0}{n}$$

характеризует относительное удлинение сообщения по сравнению с источником без избыточности (с максимальной энтропией). Здесь n — длина сообщения с энтропией $H(A)$, n_0 — минимально возможная длина сообщения с энтропией $H_{\max}(A)$.

Среднее количество информации, выдаваемое источником в единицу времени, определяется его **производительностью**:

$$H'(A) = v_{\text{и}} H(A),$$

где $v_{\text{и}}$ — скорость выдачи символов.

Производительность измеряется в битах в секунду: [бит/с].

1.2 Определение энтропии языка

Все естественные языки характеризуются достаточно большой избыточностью, которая обуславливается как неравновероятностью отдельных символов, так и наличием глубоких информационных связей между соседними символами (а также словами и целыми фразами).

Существует два подхода к определению энтропии языка, в предельном случае приводящие к одному и тому же результату [4].

С одной стороны, энтропию языка можно оценить как энтропию группы из n символов A_n, поделённую на количество символов в группе:

$$H_n^+ = \frac{H(A_n)}{n}.$$

Группа из n символов называется n**-граммой**, а величина H_n^+ — **удельной энтропией** символа.

С другой стороны, энтропию языка также можно оценить как **условную энтропию** n-го символа a_n при известных $n-1$ предыдущих символах — $H(A|A'_{n-1})$:

$$H_n^- = H(A|A'_{n-1}) = H(A_n) - H(A_{n-1}).$$

В пределе, при $n \to \infty$, обе эти величины сходятся к одному и тому же значению

$$H_{\text{яз}} = H_\infty = \lim_{n\to\infty} H_n^+ = \lim_{n\to\infty} H_n^-,$$

являющемуся энтропией данного языка [4].

В [4] также доказывается, что H_n^+ является верхней, а H_n^- — нижней границей H_∞.

1.3 Опыт Шеннона

Оригинальный способ определения энтропии языка был предложен в 1951 г. Шенноном [3]. Он заключается в отгадывании n-й буквы текста при известных $n-1$ предыдущих. Мера степени неопределённости данного опыта является оценкой сверху условной энтропии.

Из осмысленного текста наугад выбираются $n-1$ символов и кому-либо предлагается угадать n-й символ. Многократное повторение опыта даёт распределение частот правильного угадывания: частоты (вероятности) $w_1, w_2, \ldots, w_K$ того, что символ будет правильно угадан с $1, 2, \ldots, K$-й попытки (K — объём алфавита). Эти вероятности являются оценкой вероятностей символов алфавита, расположенных в порядке убывания частот [4]. Отсюда следует, что энтропия данного распределения будет являться оценкой (сверху) условной энтропии

$$H(A|A'_{n-1}) \leqslant H(W) = -\sum_{i=1}^{K} w_i \log w_i,$$

которая с увеличением n будет стремиться к энтропии языка.

Результат опыта зависит от «литературного чутья» и добросовестности отгадывающего. Для уменьшения влияния этих факторов, Шеннон предложил задавать вопросы ряду лиц и остановиться на том из них, ответы которого окажутся наиболее удачными.

1.4 Определение средней длины слова

Зная вероятность (частость) появления символа «пробел» ($\sqcup$), можно определить среднюю длину слова:

$$\overline{n}_{\text{сл}} = \lim_{N\to\infty} \frac{N - n_\sqcup}{n_{\text{сл}}} = \lim_{N\to\infty} \frac{N - NP(\sqcup)}{NP(\sqcup) + 1} = \frac{1 - P(\sqcup)}{P(\sqcup)}.$$

Здесь N — длина текста (устремлённая к бесконечности), $n_\sqcup$ — число пробелов в тексте, $n_{\text{сл}}$ — число слов в тексте, $P(\sqcup)$ — вероятность появления пробела.

Число слов в тексте на 1 больше числа пробелов: $n_{\text{сл}} = n_{\text{␣}}+1$ (если считать, что текст начинается и заканчивается буквой и все слова разделены пробелами). Число пробелов связано с длиной текста через вероятность появления пробела: $n_{\text{␣}} = NP(\text{␣})$. Вычитая число пробелов из общего числа символов, получим суммарную длину всех слов в тексте. Разделив эту величину на число слов в тексте, получим среднюю выборочную длину слова. Устремляя размер выборки (длину текста) к бесконечности, получим искомое значение $\overline{n}_{\text{сл}}$ для произвольного текста.

Для текста на русском языке (32-символьной модели) средняя длина слова примерно равна 5 символам.

2 Домашнее задание

1. Засеките время (от 1 до нескольких минут) и наберите на компьютере произвольный текст, используя только русские буквы в нижнем регистре и пробел (без знаков препинания). Можно переписать текст из книги, учебника, лекции, набрать любой текст по памяти и т.п., главное, чтобы текст был осмысленным. **Набранный текст включается в текст домашнего задания!**

2. Подсчитайте количество символов в набранном тексте. Если текст набирался в обычном текстовом редакторе в кодировке WIN или DOS, то количество символов будет равняться размеру полученного файла в байтах. Если текст набирался в визуальном редакторе (Open Office, MS Word и т.п.), то количество символов можно найти в меню статистики файла: `Файл` → `Свойства` → `Статистика`.

3. Полагая энтропию русского языка равной $H_{\text{рус}} = 1{,}37$ [бит/симв] (из [4]), вычислите количество информации I, содержащееся в набранном отрезке текста.

4. Зная количество информации, число символов и время, потраченное на набор текста, вычислите производительность источника H'.

5. Вычислите избыточность набранного текста, полагая, что объём алфавита источника $K = 32$ (на самом деле 34, если вы использовали «ё» и «ъ», но для 32-символьной модели их не учитывают, заменяя «ё» на «е» и «ъ» на «ь»).

Если возможности набирать текст на компьютере нет, можно писать текст от руки.

3 Указания к выполнению работы

Перед началом работы ознакомьтесь с интерфейсом пользователя: нажмите `F1` или выберите в главном меню «`Помощь` → `Руководство пользователя`».

3.1 Цель работы

Целью работы является анализ информационных характеристик дискретных источников.

3.2 Общие замечания

Для получения достоверных результатов, особенно при больших длинах n-грамм, требуется большой объём статистики. Кроме того, энтропия может существенно отличаться для текстов различных стилей: стихи, как правило, имеют меньшую энтропию, чем проза, энтропия делового текста меньше энтропии литературного и т.п.

Если расчёт идёт слишком медленно, **ограничьте размер исходного текста 20–30 тысячами символов**: отметьте пункт «Ограничить размер» и укажите размер текста в окошке «Макс. длина». Однако, в этом случае достоверность получаемых результатов снизится. Степень достоверности результатов можно контролировать с помощью кнопки «Статистика», отображающей таблицу распределения n-грамм. Если большая часть n-грамм встречается в тексте менее 10 раз, получаемое значение энтропии будет далеко от реального.

Объём алфавита K текста на русском языке в расчётах полагается равным 32, на английском, немецком и французском — 27 (пробел и строчные буквы без акцентов).

Имена файлов для исследований задаются преподавателем!

3.3 Снятие гистограмм распределения

Зарисуйте (сохраните в файл) гистограммы однобуквенного распределения для различных языков:

1. Откройте исходный текстовый файл на русском языке.
2. Нажмите кнопку «Гистограмма».
3. Зарисуйте (сохраните) полученную гистограмму.
4. Запишите (скопируйте) значения числа встречаемости для всех символов алфавита.

Проделайте всё то же самое для текстов на английском, немецком и французском языках.

Проанализируйте полученные распределения. Сделайте выводы.

Рассчитайте вероятность (частость) символа «пробел» $P(\sqcup)$ как отношение числа пробелов к общему числу символов в тексте и найдите среднюю длину слова $\overline{n}_{\text{сл}}$ для всех четырёх языков.

3.4 Исследование энтропии различных языков

Снимите зависимость энтропии от длины n-граммы:

1. Откройте исходный текстовый файл на русском языке.
2. Установите параметр n равным 1.
3. Нажмите кнопку «Рассчитать энтропию».
4. Занесите в табл. 1 значения длины n-граммы, рассчитанные значения энтропии n-граммы $H(A_n)$ и удельной энтропии $H(A_n)/n$.
5. Повторите п. 3–4 для значений $n = 2 \dots 5$.
6. Рассчитайте условную энтропию $H(A|A'_{n-1})$ для $n = 2 \dots 5$ как разность текущего $H(A_n)$ и предыдущего $H(A_{n-1})$ значений энтропии n-граммы.

Таблица 1

Длина n-граммы, n [симв]	Энтропия n-граммы, $H(A_n)$ [бит/симв]	Удельная энтропия, $H(A_n)/n$ [бит/симв]	Условная энтропия, $H(A\|A'_{n-1})$ [бит/симв]
1			—
...			
5			

Снимите аналогичную зависимость для текстов на английском, немецком и французском языках. Сделайте вывод о зависимости удельной и условной энтропии от длины n-граммы.

Примите за энтропию языка соответствующие значения удельной энтропии, полученные для $n = 2$. Рассчитайте избыточность для каждого языка. Сделайте выводы.

3.5 Генератор случайных текстов

Пронаблюдайте за работой генератора случайных текстов для русского языка:

1. Откройте исходный текстовый файл на русском языке.
2. Установите параметр n (длина n-граммы) равным 1.
3. Нажмите кнопку «Случайный текст». Программа сгенерирует некоторое количество символов случайного текста.
4. Увеличивая значение n до 5 с шагом 1 пронаблюдайте за изменением генерируемого текста.
5. Запишите (скопируйте) несколько «слов» случайного текста при различных значениях n.
6. Сделайте выводы.

3.6 Опыт Шеннона

Определите энтропию русского языка методом Шеннона.

1. Откройте исходный текстовый файл на русском языке.
2. Установите параметр n (длина n-граммы) равным 6.
3. Нажмите кнопку «Опыт Шеннона».
4. Попытайтесь угадать последнюю букву случайной n-граммы. Не сводите угадывание к простому перебору всех символов алфавита, используйте свои знания структуры русского языка. В трудных случаях (например, в начале слова) используйте хотя бы известное вам из п. 3.3 распределение вероятностей одиночных символов.
5. Повторите эксперимент как можно большее число раз (хотя бы 20–30, лучше 100 и больше).
6. Нажмите кнопку «Энтропия».
7. Зарисуйте (сохраните) полученную гистограмму распределения и запишите соответствующее ей значение энтропии языка.

4 Содержание отчёта

1. Название и цель работы.
2. Выполненное домашнее задание.
3. Графики, таблицы, расчётные значения и выводы по всем пунктам работы.
4. Общий вывод по результатам лабораторной работы.

Примечание: общий вывод не должен быть перефразировкой целей работы, а должен содержать обобщение (но не копию!) всех выводов, сделанных по каждому пункту в отдельности.

5 Контрольные вопросы

1. Какие источники называют дискретными? Что называется алфавитом источника?
2. Что такое «информация»? Как определяется количество информации?
3. Дайте определение энтропии дискретного источника. Как рассчитывается энтропия источника без памяти?
4. Какие источники называются источниками с памятью? Как рассчитывается энтропия источника с памятью?
5. В каком случае энтропия будет максимальной? Как определить максимальное значение энтропии?
6. Дайте определение избыточности источника. Какие источники называют безызбыточными?
7. Какую величину называют производительностью источника? В каком случае производительность будет максимальной?
8. Какую величину называют энтропией языка? Отличаются ли энтропии различных языков? Почему?
9. Как экспериментально определить энтропию языка через удельную и условную энтропию?
10. Опишите суть опыта Шеннона для определения энтропии языка.

11. Как определить среднюю длину слова для текста на заданном языке?

12. Пользуясь результатами лабораторной работы, изобразите гистограмму однобуквенного распределения для русского языка. Поясните, как по этой гистограмме определить энтропию и избыточность.

13. Пользуясь результатами лабораторной работы, изобразите зависимость условной энтропии от длины n-граммы. Поясните характер этой зависимости.

Литература

1. Кловский Д. Д. Теория электрической связи. — М.: Радиотехника, 2009. — 648 с.

2. Теория электрической связи: учебник для вузов / А. Г. Зюко, Д. Д. Кловский, В. И. Коржик, М. В. Назаров; Под ред. Д. Д. Кловского. — М.: Радио и связь, 1998. — 432 с.

3. Шеннон К. Работы по теории информации и кибернетике. — М,: Издательство иностранной литературы, 1963.

4. Яглом И. М., Яглом А. М. Вероятность и информация. — М.: Наука, 1973.